

Amendments to the Specification:

Please replace paragraph [0009] with the following amended paragraph:

[0009] There have been several attempts to address this problem by creating search tools, such as MedLine, MEDLINE[®], Chemical Abstracts[®], Biosis-Previews, BIOSIS Previews[®], etc., that permit computer searching of large numbers of scientific journals or abstracts, such as Science, Nature, Proceedings of the National Academy of Sciences, etc. Searching these journals is still a problem because there are hundreds of such journals and many can only be searched by key words (and searching is sometimes restricted to key word fields or abstracts) or by reading full abstracts, which in either case is very time-consuming and inefficient such that important articles are easily missed.

Please replace paragraph [0010] with the following amended paragraph:

[0010] Another partial solution is databases of genomics data. One example is GenBank[®], which is maintained by the National Center for Biotechnology Information (NCBI). Gene sequences entered in such databases are usually annotated with information that may include, e.g., the type of cell in which a given gene sequence is expressed, the probable function of the sequence, etc.

Please replace paragraph [0074] with the following amended paragraph:

[0074] Fig. 3 shows the conceptual components of the analysis. The data structures, algorithms, and software components used to perform the analysis may form a stand alone software tool or they may be integrated with an existing platform and/or suite of applications that are used to access information stored in the KRS. The analysis may include two steps. A first step involves a series of computations over a copy of the KB to identify profiles, and a second step that involves scoring these profiles against user provided data. In the following description and in reference to Fig. 3, an example of the analysis uses user-supplied expression array data. A library (7) of profiles is preferably generated according to a user data set, e.g., user-supplied differential gene expression data, but in other embodiments profiles may be pre-generated independently of the user data. The nature of the generated profiles may vary considerably based on the goals of the analysis, as is explained in greater detail below. In an alternative embodiment, a pre-generated "library" of

profiles, mapping an entire KB, may be preferred for the sake of performance — pre-generate all of these maps so that retrieving them later will be faster. The user-supplied data may include array data provided from a third party product, e.g., an Affymetrix GeneChip^{[[c)]]}, online service or proprietary database.

Please replace paragraph [0090] with the following amended paragraph:

[0090] Results from the analysis may be presented to the user in various forms. In one embodiment, three types are presented:

1. The first is a list of profiles ranked according to a profile score (14), generated by calculating the P-value for each profile (13) in the library and sorting the resulting list. Each profile lists the gene central to the profile, and any genes from the expression dataset that also appear in the profile. Users can view this list and pick profiles that appear to be interesting to look at them in greater detail. This output may be viewed using a spreadsheet program.
2. The second is one or more profile diagrams (17) for each of the profiles. These diagrams show all the genes from the profile and the key relationships between them in the form of a “circles and arrows” diagram. Different symbols, colors, labels, and positions are used to encode additional information about the profile which is extracted (16) from the KB. Different diagrammatic representations may be used to display the same underlying profile but highlighting different characteristics. For example, one diagram may use a layout algorithm that highlights the subcellular localization of the gene product by grouping symbols together if they share the same localization. Another diagram may use a layout algorithm that highlights the interrelationships between gene products by grouping symbols together if they share many interactions. An example of such information is the subcellular localization of the gene product (information that can be stored in the KB but is not used for profile generation or profile scoring in a preferred embodiment). The diagram itself may be generated (15) using an open source/freely-available 3rd party diagramming tool from AT&T Research called GraphViz (see www.graphviz.org). The output may be a printout of a diagram or a web-accessible graphic (image file or Scalable Vector Graphics — SVG-file).

3. The third is algorithmic association of biological processes with pathway profiles (18). This step involves generating a description or summary of the biology manifested by a given profile by performing algorithmic analyses of the findings relating the genes in the profile. Conceptually this is analogous to automatically generating a set of labels or captions (18) that describe the molecular, cellular, organismal and/or disease processes that best represent the function(s) of the genes in this profile. For example, while many cellular processes may be involved in the various genes in the profile, “apoptosis” may stand out as statistically significant among them. Inferred processes can be derived from the findings to collect findings that support the involvement of genes in more general processes using the ontology hierarchy. For instance, some genes may ‘increase apoptosis’ and others may affect ‘apoptosis of T cells’, yet all of those genes can be inferred to be involved in ‘apoptosis.’ This aspect of results creation is particularly powerful since it leverages the unique structure of an ontology. These process annotations — e.g., the most representative or highest scoring ones — may appear on the diagram itself, or may be supported by a more complete list on a separate page, or via a web display that supports iterative “drill down” to reveal additional details. The output may be a text printout, but may also be presented to the user in a GUI interactive form. Specifically, the findings in the KB structure information about processes such that, for example, the process, the location(s) in which the process occurs, when the process occurs, the molecule(s) that initiated or affected the process, and the objects acted on by the process are distinguished from each other. The association between molecules and the processes they are involved in can be constructed by first building a graph (tree) of processes, starting with nodes that represent the detailed processes (e.g., ‘increases arrest in G2 phase of fibroblasts’), and then deriving (more general) parent processes by successively removing (e.g., ‘increases’) or generalizing (e.g., ‘fibroblasts’ are ‘cells’ based on relationships in the ontology) details of those process. These generalized processes are not necessarily stated explicitly in any findings, but instead are inferred from the specific processes that are stated explicitly in the findings and inference rules based on relationships in the ontology. Thus, the presence of an ontology allows a single stated set of relationships (the finding) to imply a much larger set of relationships that can still be used for computation and display to users. After the process tree is constructed, genes and the findings that

support the detailed processes are also relevant to their parent processes so they are inferred up the process tree. Therefore, very general processes at the top of the tree (e.g., 'apoptosis') may be associated with all the genes and findings for all their more-specific child processes (e.g., apoptosis of specific cells, directions of effect on apoptosis, etc.). Thus, the process tree aggregates information at different levels of detail, from specific to general, and the molecules associated with each process annotation are compared to the molecules in each profile to score them.